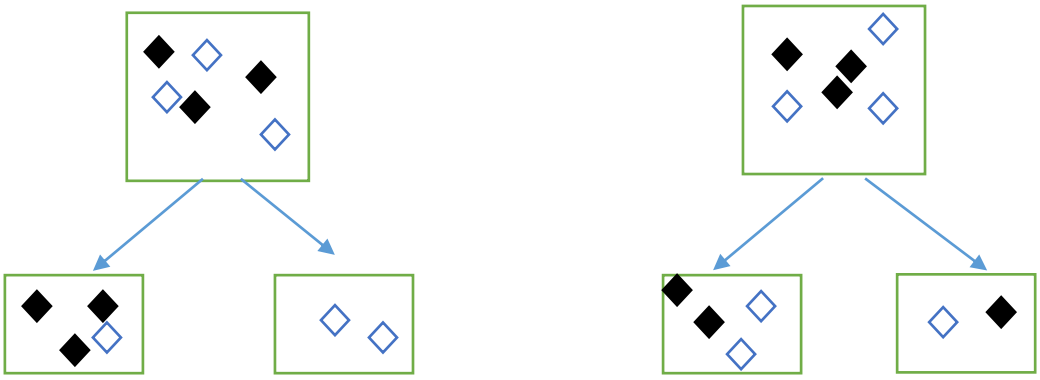
1. Consider the possible splits of training records during the decision tree induction.



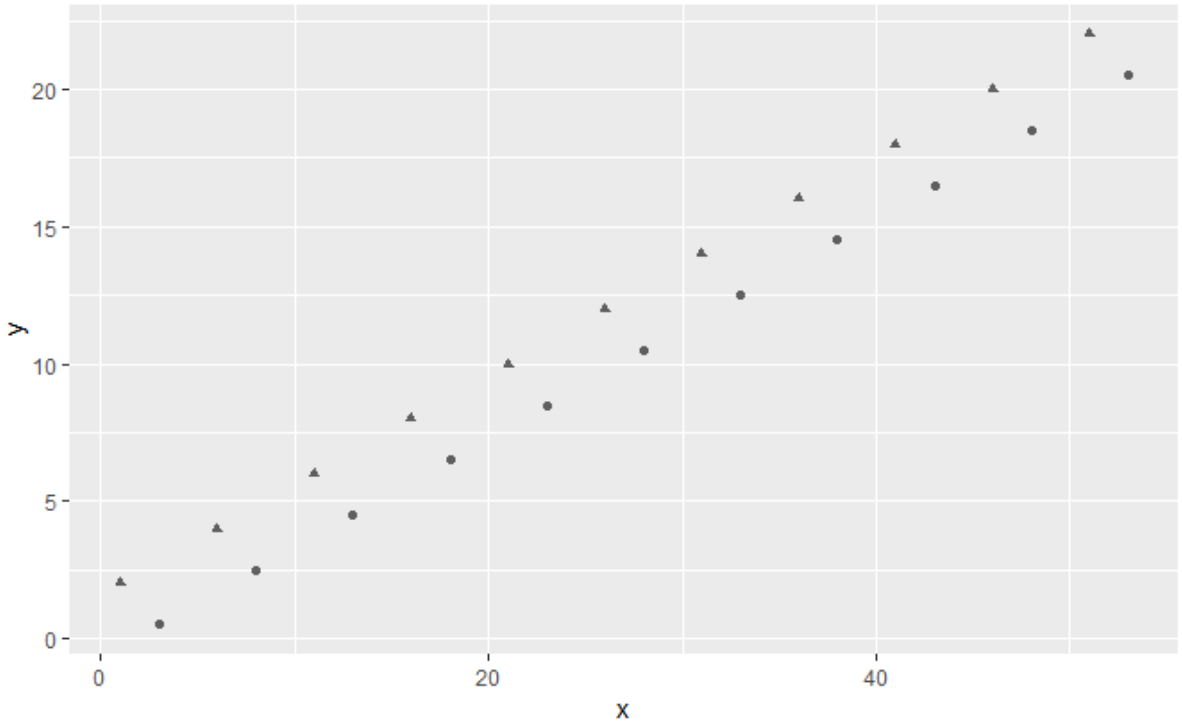
Possible splits of training records according to attributes x1 and x2. Black and white colors represent class labels.

What is the Gini index of each split? Which split produces more pure nodes?

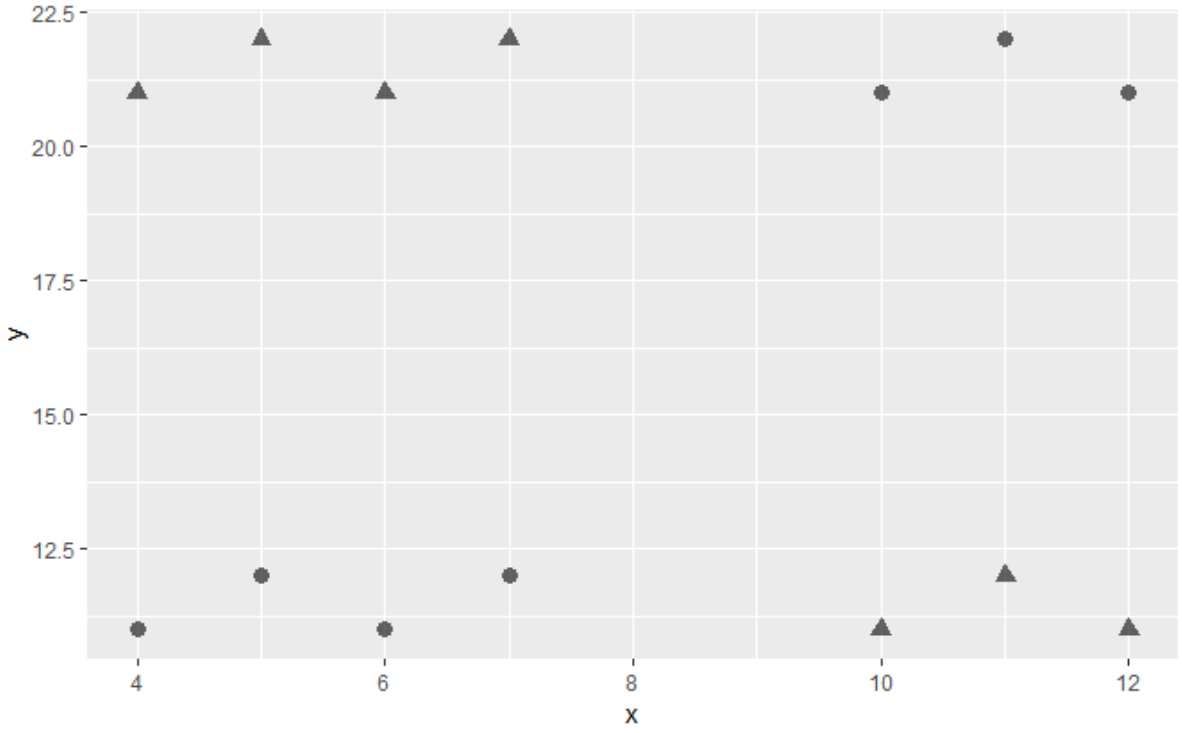
2. Given the following data. use KNN regression to calculate the predicted value of target variable, when k equals to the number of observations. Predict the value of y, when x = 5.

x	y
-1.6	50
-1.4	70
NA	70
-0.8	100
-0.5	115
NA	130
-0.3	100
-0.2	70
0.4	155
0.5	201
0.6	250

3. Suppose we have the following data represented using two real-valued features (X and Y) and suppose that our goal is to randomly split this data into a training set (90%) and a test set (10%) and to train and evaluate a model. Does the KNN (with k=1) have any chance of doing well in terms of accuracy?



4. Suppose we have the following data represented using two real-valued features (X and Y) and suppose that our goal is to randomly split this data into a training set (11 points) and a test set (3 points) and to train and evaluate a model. Does the KNN (with k=1) have any chance of doing well in terms of accuracy?



5. Perform 2-means clustering, using the data from the table with 2 features:

var1	var2
2	5
8	3
0	4
0	5
7	3

Randomly (or not) assign two centroids to start the calculations. Assign each observation to the centroid to which it is closest, in terms of Manhattan distance. Report the cluster labels for each observation. Repeat the steps until your centroids stop changing. What is the final cluster membership? How many iterations have you used?

6. Use single link agglomerative clustering to group the data described by the following distance matrix. Show the dendograms.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0